

Mit Big Data gegen das "Ende der Theorie"?

Mahrt, Merja

Erstveröffentlichung / Primary Publication

Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:

Mahrt, M. (2015). Mit Big Data gegen das "Ende der Theorie"? In A. Maireder, J. Ausserhofer, C. Schumann, & M. Taddicken (Hrsg.), *Digitale Methoden in der Kommunikationswissenschaft* (S. 23-37). Berlin <https://doi.org/10.17174/dcr.v2.2>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
<https://creativecommons.org/licenses/by/4.0/deed.de>

Terms of use:

This document is made available under a CC BY Licence (Attribution). For more information see:
<https://creativecommons.org/licenses/by/4.0>

Empfohlene Zitierung: Mahrt, M. (2015). Mit Big Data gegen das „Ende der Theorie“? In A. Maireder, J. Ausserhofer, C. Schumann, & M. Taddicken (Hrsg.), *Digitale Methoden in der Kommunikationswissenschaft* (S. 23-37). doi: 10.17174/dcr.v2.2

Zusammenfassung: Die Analyse großer Datensätze aus Online-Plattformen („Big Data“) wirft eine Reihe von methodologischen Fragen auf, die insbesondere das Verhältnis von Theorie und Empirie betreffen. Während an vielen Stellen Begeisterung für neue, datenorientierte Forschungsmöglichkeiten zu hören ist, befürchten andere ein „Ende der Theorie“. Der Beitrag zeigt verschiedene Aspekte dieses Problembereichs auf und macht Vorschläge, wie Big-Data-Studien theoretisch gehaltvoll gestaltet werden können. Beispiele aus der Agenda-Setting-, Sozialkapital- und Diffusionsforschung werden diskutiert. Interdisziplinäre Zusammenarbeit und systematische Validierung von Big-Data-Verfahren erscheinen dabei als zentrale Desiderate. Welchen Stellenwert Big-Data-Ansätze in der kommunikationswissenschaftlichen Forschung und Lehre in Zukunft haben werden, hängt mutmaßlich auch von ihrer Fähigkeit zu theoretischer Innovation ab.

Lizenz: Creative Commons Attribution 4.0 (CC-BY 4.0)

Merja Mahrt

Mit Big Data gegen das „Ende der Theorie“?

In den letzten Jahren hat sich eine Reihe kommunikationswissenschaftlicher Zeitschriften und Tagungen mit „Big Data“ befasst. Die Verfügbarkeit großer Mengen an Daten aus Online-Plattformen zur Informationssuche, Rezeption medialer Inhalte und interpersonalen Kommunikation hat neue Arten der Analyse hervorgebracht, die allerdings im Fach nicht unumstritten sind. Der vorliegende Beitrag beleuchtet vor allem die Diskussion um den theoretischen Gehalt von solchen oft eher datengetriebenen Untersuchungen und versucht aufzuzeigen, wie Big-Data-Verfahren auch einen theoretischen Mehrwert erbringen können. In einem ersten Schritt werden Hoffnungen zusammengefasst, die sich mit Big Data verbinden. Das zweite Kapitel beschreibt drei Problembereiche, die sich aus typischen Vorgehensweisen in Big-Data-Untersuchungen ergeben. Zu diesen werden daraufhin Lösungen vorgeschlagen und abschließend einige Annahmen zur Zukunft von Big-Data-Forschung in der Kommunikationswissenschaft aufgestellt.

1 Euphorie für Big Data

Mit dem Schlagwort Big Data verbinden sich verschiedene Formen der Analyse „großer“ Datensätze. Im ursprünglichen, informatischen Sinn handelt es sich dabei um Datenmengen, die zu groß sind, um sie mit üblicher Hard- und Software zu verarbeiten (Manovich, 2012). Die absolute Größe von Daten, die als

Big Data bezeichnet werden, wächst dadurch im Laufe der Zeit. Für die Sozialwissenschaften verbindet sich mit dem Begriff in der Regel ein bestimmter Typ von Datensatz, der nach informatischen Maßstäben nicht unbedingt groß sein muss, aber dennoch deutlich umfangreicher sein kann als etwa übliche Befragungs- oder inhaltsanalytische Datensätze.

Vor allem geht es in der sozial- und speziell kommunikationswissenschaftlichen Diskussion um den Begriff aber um die neuartige Herkunft und Struktur dieser Daten. Während klassischerweise Daten mittels eines im Forschungsprozess entwickelten Messinstruments (Interviewleitfaden, Fragebogen, Codebuch, Beobachtungsschema...) erfasst werden, nutzen viele Big-Data-Studien Daten, die nicht speziell zum Zweck der Analyse entstanden sind. Sie fallen etwa als Spuren der Nutzung digitaler Technologien und Medien an. Häufig unterscheidet sich daher ihre Fall- und/oder Variablenstruktur von typischen sozialwissenschaftlichen Datensätzen. Es können für einen Nutzungsvorgang zum Beispiel sehr viele Datenpunkte in kurzem Abstand erfasst werden, die einen Zeitverlauf sehr engmaschig dokumentieren. Wenn es sich um Kommunikationsdaten handelt, ergeben sich häufig außerdem Kontakt- oder Netzwerkstrukturen. Beide Arten von Strukturen lassen sich mit klassischen Erhebungsverfahren nur sehr aufwändig in ähnlicher Detailgenauigkeit oder Menge erfassen. Auf der anderen Seite sind in Big-Data-Studien häufig die Informationen nicht verfügbar, die man mit üblichen Verfahren als unabhängige oder Kontrollvariablen über die Personen erheben würde, deren Verhalten untersucht wird. Groß sind Big-Data-Datensätze in diesen Fällen dann nur in einer Dimension der Datenmatrix: Es können zwar sehr viele Fälle, aber oft nur wenige Variablen, also Informationen pro Fall, untersucht werden.

Während etwa Telefongesellschaften oder Einzelhandelsunternehmen durch den Einsatz elektronischer Verfahren schon seit längerem Kommunikations- bzw. Konsumverhalten detailliert erfassen und analysieren können (Savage & Burrows, 2007), hat die sozialwissenschaftliche Diskussion um ähnliche Ansätze etwas später eingesetzt. Sie kam insbesondere mit der steigenden Nutzung von Online-Plattformen auf, die Spuren von Nutzerverhalten entweder öffentlich machen oder über die diese Spuren durch Schnittstellen auslesbar werden. Von der angewandten Markt- und Meinungsforschung werden diese Möglichkeiten nach wie vor als große Verheißung angepriesen, die zum Beispiel ein besseres, schnelleres oder umfassenderes Verständnis von Kundenwünschen und -verhalten erlaubten (z. B. Kearon & Harrison, 2011; Russom, 2011).

Dagegen ist die Reaktion der Sozialwissenschaften vielschichtiger: Auf der einen Seite lassen sich auch hier Euphorie für neue Forschungsmöglichkeiten und Datenstrukturen verzeichnen (Lazer et al., 2009; Miller, 2011). Auf Tagungen und in jüngeren Publikationen finden sich in den letzten Jahren so immer wieder Untersuchungen, die mit großen Datensätzen aus Online-Plattformen arbeiten, etwa Twitter-Meldungen, Google-Suchanfragen oder Anzahl von „Freunden“ auf Facebook. Auf der anderen Seite hat sich aber auch eine kritische Diskussion um Big-Data-Forschung dieser Art entwickelt. Diese beleuchtet etwa methodische Unterschiede zwischen traditioneller Sozialforschung und Big-Data-Ansätzen und Implikationen, die sich hieraus für die Aussagekraft der gewonnenen Befunde ergeben (boyd & Crawford, 2012; Couper, 2013; Mahrt & Scharkow, 2013; Manovich, 2012; Tinati, Halford, Carr, & Pope, 2014). Auch rücken neue ethische Probleme in den Vordergrund, die etwa die fehlende Einwilligung der Nutzer, deren Verhalten untersucht wird, oder den Schutz privater Daten betreffen (siehe Heise in diesem Band; Markham & Buchanan, 2012; Zimmer, 2010). Neben solchen eher forschungspraktischen Aspekten werden in Bezug auf Big Data aber vor allem theoretische Fragen diskutiert. In verschiedenen Variationen ist hierbei von einem „Ende der Theorie“ die Rede.

2 Big Data und das „Ende der Theorie“

Als Theorien werden in den Sozialwissenschaften streng genommen in sich widerspruchsfreie Systeme von Aussagen verstanden, aus denen sich Hypothesen ableiten lassen (Schnell, Hill, & Esser, 2005). Auf dieses Verständnis von Theorie bezieht sich auch die Diskussion um deren (vermeintliches) Ende durch Big Data. In einer positiven Sichtweise machen die neuen großen Datensätze die Beschäftigung mit Theorien überflüssig oder stellen zumindest ihre traditionelle, hypothesengenerierende und forschungsleitende Rolle in Frage. Anderson (2008) etwa beschreibt, wie mit ausreichend großen Mengen an Daten auch Korrelationen für Wissenszuwachs sorgen können, anstatt aus Theorien Annahmen abzuleiten und anschließend entsprechende Kausalmodelle zu prüfen. Durch die Fülle an Daten und heute verfügbare Rechnerkapazität, diese zu analysieren, werde der übliche Prozess der Aufstellung und Überprüfung von Hypothesen durch letztlich aussagekräftigere datengetriebene Befunde ersetzt. Während theoretisch abgelei-

tete Modelle immer vereinfachen und dadurch auch verfälschen, könnten nach Anderson schier unüberschaubare Mengen an Daten die Notwendigkeit solcher Modelle (die er als wenig mehr als Hilfskonstruktionen darstellt) überflüssig machen. Echte – im Sinne von: nicht durch ein Modell um ihre natürliche Komplexität reduzierte – und in unbeschränkter Menge verfügbare Daten sieht er daher gegenüber traditionell erhobenen im Vorteil.

Ein konkretes Beispiel soll diese Vorgehensweise illustrieren: In seinem Beitrag zur Abschlussdiskussion der ICA-Konferenz in Phoenix 2012 hat Bailenson (2012) eine entsprechende Studie aus der Kommunikationswissenschaft vorgestellt. Um das Verhalten von Lehrenden im Klassenraum und die Reaktionen von Schülerinnen und Schülern darauf zu untersuchen, hat er Sensorleisten aus Computerspielen eingesetzt, die Körperhaltungen und Bewegungen im Raum aufzeichnen können. Die resultierenden Daten sind in Zeitverlauf und Detailreichtum wesentlich umfangreicher als bei klassischen Beobachtungsverfahren. Dadurch konnte der Forscher Zusammenhänge zwischen Lehrer- und Schülerverhalten entdecken, die anderweitig nicht erfassbar gewesen wären. Es ergaben sich Korrelationen zwischen minimalen Unterschieden in Haltung und Bewegung der Lehrenden sowie der Reaktionen der Schülerinnen und Schüler. Bailenson suchte also tatsächlich in riesigen Datenmengen nach Mustern und Zusammenhängen, die in einer kleineren Untersuchung entweder nicht erfasst oder im Rauschen von Messfehlern und statistischer Irrtumswahrscheinlichkeit übersehen worden wären. Provokativ spitzte er zu, dass Theorie im Forschungsprozess damit überflüssig sei, denn auf Basis des vorhandenen Wissens hätte er niemals zu Hypothesen über die sich als relevant erweisenden Verhaltensaspekte und ohne den Big-Data-Ansatz auch nicht zu vergleichbaren Befunden gelangen können.

In seiner Keynote zur „Digital Methods“-Tagung hat Jürgen Pfeffer (2013) ein solches Vorgehen als typisch für die Arbeit mit Big Data beschrieben: Das Prinzip des klassischen Forschungsprozesses (von ihm in folgender Reihenfolge skizziert: Problem, Fragestellung/Hypothesen, Theorie, Methode, Daten, Analyse, Ergebnispräsentation) sei quasi auf den Kopf gestellt, weil erstens mit der Methode begonnen werde, woraus sich zweitens Daten ergäben, die drittens einer Analyse unterzogen würden, deren Ergebnisse dann viertens präsentiert würden. Und erst daran schließe sich fünftens die Darstellung eines Problems an, auf das die Untersuchung eine Antwort gegeben habe. Wohlgermerkt geht es in Pfeffers Worten hier um ein gesellschaftliches oder anderes *Problem*. Ein

Bezug zu *Theorie* im eigentlichen Sinn wird in solchen Forschungsarbeiten nicht notwendigerweise hergestellt.

Es mag genügend gesellschaftliche Probleme geben, über die man mit einem solchen Vorgehen Befunde liefern kann. Und nicht jede Untersuchung muss zur Theorieentwicklung beitragen wollen. Ein Vorgehen wie von Pfeffer zusammengefasst schließt aber theoretische Fortschritte tatsächlich aus, weil es ohne jeglichen Bezug zu Theorie auskommt. Im Gegensatz zum induktiven Vorgehen im herkömmlichen sozialwissenschaftlichen Verständnis wird an die Auswertung der Daten eben keine Theoriediskussion angeschlossen. Dieses erste „Ende der Theorie“ wird wie oben beschrieben von einigen begrüßt und kann durchaus zu neuen Erkenntnissen führen. Im eigentlichen Sinn wird Theorie aber nicht weiterentwickelt, weil man sich bei dieser Variante des Forschungsprozesses nicht mit übergeordneten widerspruchsfreien Systemen von Aussagen beschäftigt und keine Gesetzmäßigkeiten oder Regelmäßigkeiten jenseits des untersuchten Problems ableiten möchte.

Daneben existiert noch eine zweite Variante des „Endes der Theorie“, die sich zuweilen in Big-Data-Studien zeigt. Dies ist dann der Fall, wenn es zwar dem Wortlaut nach in einer Untersuchung Bezüge zu bestehenden Theorien gibt, diese aber durch die methodische Umsetzung nicht wirklich eingelöst werden können. Insbesondere drei Bereiche erscheinen hierbei problematisch.

(1) Aussagekraft von Big-Data-Stichproben

Big-Data-Analysen mit online abrufbaren Daten sind häufig auf ad-hoc-Stichproben angewiesen, deren Repräsentativität und Vergleichbarkeit mit anderen Stichproben unklar ist. Zum einen gibt es in der Regel keine Möglichkeit, Verteilungen der Stichprobe mit bekannten Parametern der Grundgesamtheit zu vergleichen. Zum anderen sind die genauen Bedingungen der Datenerhebung häufig nicht reproduzierbar und können daher nur schwer überprüft werden (Gaffney & Puschmann, 2014). Dies schränkt die Aussagekraft der daraus abgeleiteten Befunde ein, etwa in einer Agenda-Setting-Studie mit Twitter-Daten: Um die Erfassung einer Publikumsagenda haben sich in der Agenda-Setting-Forschung eine umfassende Diskussion und darauf aufbauend verschiedene Standardbefragungsinstrumente entwickelt (Dearing & Rogers, 1996). Ein Big-Data-Ansatz kann dagegen darin bestehen, Twitter-Meldungen für die Messung der Wichtigkeit von Themen

in der Bevölkerung zu verwenden (z. B. Verdegem & D'heer, 2013). Je nach Art des Zugriffs auf die Daten gibt es aber mehrere Stellen, an denen die Plattform Twitter möglicherweise bereits eine Auswahl aus allen Meldungen trifft, die das Aufgreifkriterium erfüllen. Dies ist in der Regel für Forschende nicht transparent (Gaffney & Puschmann, 2014).

Ein weiteres Stichprobenproblem ergibt sich aus der Frage, wofür die Twitter-Meldungen stehen, wen oder was sie repräsentieren sollen. Welche Begriffe oder Personen auf Twitter im Zeitverlauf wie häufig erwähnt werden, kann in der Tat Aufschluss über deren sich verändernde Wichtigkeit geben. Aber dies betrifft nur Twitter-Nutzerinnen und -Nutzer, die keinen repräsentativen Ausschnitt der Bevölkerung bilden (Busemann & Gscheidle, 2012). Eine Liste von Twitter-Themen ist damit nicht ohne Weiteres mit einer über repräsentative Befragung erhobenen Publikumsagenda vergleichbar. Selbst wenn Begriffe aus etablierten Theorien verwendet werden (wie Publikumsagenda), sind die Ergebnisse von Big-Data-Studien damit nicht unbedingt an die traditioneller Untersuchungen anschlussfähig. Die Stichproben sind zwar häufig groß, müssen aber nicht zwingend auch aussagekräftig sein. Dadurch wird der theoretische Bezug nur ein scheinbarer, und es wäre möglicherweise besser, die Daten schlicht in ihrem ursprünglichen Sinn zu beschreiben (bspw. als Themenkarriere in Twitter), anstatt sie mit theoretischer Bedeutung zu überfrachten, die sie nicht einlösen können (weil Twitter-Trends eben nicht als Indikator für breitere Themeninteressen der Bevölkerung stehen können; siehe auch Mitchell & Hitlin, 2013).

(2) Messprobleme in Big-Data-Untersuchungen

Ein zweites Problem kann entstehen, wenn Maßzahlen, die aus Online-Plattformen einfach zu generieren sind, eins zu eins auf komplexe theoretische Konstrukte übertragen werden. Beispielsweise kann man in sozialen Netzwerkseiten (etwa Facebook) in der Regel die Anzahl der Kontakte einer Nutzerin oder eines Nutzers einsehen. Dies als alleinigen Indikator für das Sozialkapital einer Person anzunehmen (wie z. B. bei Brooks, Welser, Hogan, & Titsworth, 2011), wird diesem Begriff allerdings wohl kaum gerecht, weder theoretisch (Adler & Kwon, 2002), noch in Anbetracht der Befunde zu Sozialkapital und Facebook-Beziehungen (Burke, Kraut, & Marlow, 2011; Yoder & Stutzmann, 2011). Eine Studie, die Kontakte in sozialen Netzwerkseiten mit Sozialkapital gleichsetzt, lässt sich damit nur

schwer mit dem Forschungsstand zu Sozialkapital in Beziehung setzen. Ähnliche Probleme haben alle Big-Data-Studien, die auf Daten beschränkt sind, die von Online-Plattformen vorstrukturiert worden sind. Die Art und Weise, wie Eigenschaften oder Verhalten von einer Plattform erfasst werden, lässt sich nur bedingt auf das Theorieinventar der Sozialwissenschaften übertragen und ist nicht mit der üblichen Vorgehensweise sozialwissenschaftlicher Operationalisierung von Konstrukten vergleichbar. Die Vorstrukturierung in Kategorien durch eine Plattform ist aber Voraussetzung dafür, dass große Mengen an Daten ausgelesen werden können. Wenn die Bedeutung dieser somit vorgegebenen Kategorien aber unklar ist, bleibt auch hier offen, was diese Daten jenseits des rein deskriptiven Werts eigentlich bedeuten – insbesondere für die Theorie, in deren Kontext sie erhoben werden sollen (boyd & Crawford, 2012; Mahrt & Scharkow, 2013).

(3) Fehlende theoretische Bezüge in interdisziplinären Forschungsfeldern

Ein dritter Problembereich ist nicht auf Big-Data-Forschung beschränkt, sondern entsteht dort häufig durch den besonders stark ausgeprägten interdisziplinären Charakter des Forschungsfeldes Online-Kommunikation. Stellvertretend für viele andere Bereiche kann man das am Beispiel der Diffusionsforschung illustrieren. Eine viel zitierte Studie hat untersucht, welche Rolle die Nähe zwischen Facebook-Nutzerinnen und -Nutzern beim Weiterleiten von Informationen innerhalb der Plattform spielt (Bakshy, Rosenn, Marlow, & Adamic, 2012). Mit 253 Millionen Probanden ist hierzu ein wahrlich großes Experiment durchgeführt worden. Der Titel der Studie, „The role of social networks in information diffusion“, klingt aus kommunikationswissenschaftlicher Perspektive zunächst vertraut; schaut man sich die in der Studie zitierte Literatur an, so fällt dagegen auf, dass hier ganz andere Quellen zitiert werden, als man erwarten würde. Mit Greenberg und Granovetter kommen zwar am Rande auch für dieses Fach einschlägige Forscher vor, der Fokus der theoretischen Auseinandersetzung liegt bei Bakshy et al. aber anderswo. Dies ist ihr gutes Recht und nicht der Kern des Problems. Es geht hier nicht um den Vorwurf, Bakshy oder andere hätten aus der Perspektive der Kommunikationswissenschaft zentrale Studien nicht gelesen. Ebenso kann kein Fach den einzigen (oder einzig richtigen) Zugang zu einem Forschungsgegenstand für sich beanspruchen. Es fällt aber auf, dass mit der (datenorientierten) Online-Forschung wiederum ein Bereich entsteht, in dem kommunikationswissenschaft-

liche Forschung mehr oder weniger unverbunden neben Arbeiten aus anderen Fächern steht, und das selbst in Fällen wie der Diffusionsforschung, in denen ein Gegenstand in zwei Fächern ähnlich untersucht wird.

Wie oben angedeutet, ist dieses Phänomen nicht neu und nicht auf Big-Data-Forschung beschränkt. In der Medizin sind ab den 1980er Jahren unabhängig von der damals bereits existierenden kommunikationswissenschaftlichen Forschung Skalen zur Fernsehnutzung sowie ihren Motiven entwickelt worden (Bickham, 2009). Umgekehrt wird ein Großteil der gut 2.000 in der medizinischen Datenbank *PubMed* verzeichneten Studien zu Fernsehnutzung in unserem Fach eher selten rezipiert. Psychologische Darstellungen zu Einstellungen, kognitiver Dissonanz und selektiver Zuwendung kommen ohne Verweise auf kommunikationswissenschaftliche Forschung aus, während in der Psychologie geläufige Dimensionen von Einstellungen in der Kommunikationswissenschaft nicht berücksichtigt werden (Mahrt, 2014). In solchen Fällen geht Potenzial verloren, Theorie (und auch Empirie) in den betreffenden Fächern weiterzuentwickeln, weil nicht berücksichtigt wird, was in nahen oder auch entfernteren Nachbardisziplinen erarbeitet wird. Dies bedeutet zwar für keines der betreffenden Fächer das Ende der (eigenen) Theorie, aber es werden doch Möglichkeiten zur interdisziplinären theoretischen Weiterentwicklung verschenkt. Dies muss sich in der Big-Data-Forschung nicht im gleichen Maße wiederholen wie in den beiden genannten älteren Forschungsbereichen, und es wäre wünschenswert, dass es nicht so kommt. Die Verfügbarkeit von disziplinübergreifenden Suchmaschinen und Katalogen macht es heute einfacher, auf fachfremde, aber inhaltsverwandte Forschung aufmerksam zu werden. Unter anderem auf diesem Weg könnte auch Big-Data-Forschung zu Online-Kommunikation integrierte Theoriebildung vorantreiben und disziplinäre Parallelentwicklungen vermeiden, wozu im folgenden Abschnitt weitere Anregungen für die drei dargestellten Problembereiche gegeben werden sollen.

3 Theoretische Fortschritte durch Big-Data-Forschung

Für die drei im vorigen Abschnitt skizzierten Problem- und Forschungsfelder sind zum Teil bereits Lösungsansätze erarbeitet worden. In anderen Fällen könnten Arbeiten jenseits der Big-Data-Forschung im hier diskutierten Sinne Anregungen für theoretisch gehaltvollere Forschung sein.

(1) Vergleichbarkeit von Big-Data-Stichproben

Wo Befunde aus Big-Data-Untersuchungen bis dato nicht mit dem Forschungsstand verglichen werden können, bietet es sich natürlich an, Studien zur Vergleichbarkeit durchzuführen. Ob sich Themenkarrieren in Social-Media-Plattformen mit üblichen Verfahren der Erfassung von Agenden vergleichen lassen, ist ja letztlich eine empirische Frage. Groshek und Groshek (2013) beispielsweise vergleichen die Beliebtheit von Nachrichtenthemen auf Twitter und Facebook mit Nachrichten des TV-Senders CNN und der New York Times. Für Facebook ergeben sich deutliche und signifikante Rangkorrelationen, für Twitter dagegen nicht. Auch zwischen der Häufigkeit von Google-Suchanfragen zu einem politischen Thema und der Wichtigkeit dieses Themas in repräsentativen Befragungen desselben Zeitraums gibt es einen Zusammenhang, dieser fällt aber eher schwach aus (Scharnow & Vogelgesang, 2011). Weitere Studien dieser Art werden es hoffentlich in Zukunft ermöglichen, die Aussagekraft von Big-Data-Stichproben und damit auch ihrer Ergebnisse besser einzuschätzen. Es wäre sogar denkbar zu prüfen, ob sich zum Beispiel für bestimmte Substichproben einer Befragung (etwa alle Twitter-Nutzerinnen und -Nutzer) ähnliche Befunde zeigen wie in den Daten, die aus der betreffenden Plattform in großer Zahl ausgelesen wurden. Die bislang vorliegenden vergleichenden Studien verdeutlichen aber auch, dass Nutzungsdaten aus Onlineplattformen eben nicht ohne weiteres für etwas anderes stehen können.

(2) Vergleichbarkeit von Big-Data-Maßzahlen

Die im vorigen Abschnitt vorgestellten Arbeiten erlauben es nicht nur, die Übertragbarkeit von Stichprobenergebnissen zu untersuchen. Auch die von einer Plattform bereitgestellten Kategorien (beliebte Begriffe einer Social-Media-Plattform, Suchanfrage bei Google) können dabei mit üblichen inhaltsanalytischen bzw. Befragungskategorien verglichen werden. Darüber hinaus sind auch systematische Validierungen von Messverfahren möglich und werden für die Big-Data-Forschung eingefordert (boyd & Crawford, 2012; Lazer, Kennedy, King, & Vespignani, 2014). In Abschnitt 2 wurde beispielsweise bezweifelt, dass die Anzahl von Kontakten auf Facebook ein valider Indikator für Sozialkapital sei. Auch dies ließe sich empirisch prüfen. Vorbild dafür könnte zum Beispiel Williams' (2006) Studie sein, in der verschiedene Maße für Online- und Offline-Sozialkapital

entwickelt und mit anderen, etablierten Indikatoren verglichen wurden. Ohne solche Validierungen bleibt letztlich unklar, was die Kategorien aus Online-Plattformen eigentlich erfassen.

(3) Theoriegeleitete Big-Data-Studien

Neben vergleichenden Studien ist es auch denkbar, die Vorteile von Big Data für die Weiterentwicklung von Theorien fruchtbar zu machen. Vorbilder dafür können Simulationsstudien sein. Watts und Dodds (2007) beispielsweise prüfen mit großen, aber simulierten Datensätzen theoretische Annahmen und Modelle der Diffusionsforschung. Sie können zeigen, wie unwahrscheinlich ein reiner Zwei-Stufen-Fluss in sozialen Netzwerken ist, bei dem Informationen aus Massenmedien zunächst nur Meinungsführer erreichen, die diese dann an ihre Kontakte (Follower) weitergeben. Wenn man wie Bakshy et al. (2012) privilegierten Zugang zu Facebook-Nutzerdaten hat, wäre es denkbar, in der immensen Stichprobe von 253 Millionen Nutzerkonten ebenfalls nach solchen Mustern zu suchen – in diesem Fall sogar unter natürlichen, alltäglichen Bedingungen. Ob Facebook-Netzwerke in Bezug auf Diffusion mit anderen sozialen Netzwerken vergleichbar sind, lässt sich damit möglicherweise auch einschätzen. Das anders gelagerte Forschungsinteresse von Bakshy et al. ist natürlich auch ohne diese Erweiterung legitim – aber die verschiedenen Fächer, die sich mit ähnlichen Big-Data-Datenquellen oder sogar ähnlichen theoretischen Konstrukten beschäftigen, könnten sich sicherlich stärker befruchten, als es zur Zeit der Fall zu sein scheint – und zwar gegenseitig.

4 Ausblick: Vorschläge für den zukünftigen Umgang mit Big Data

Zusammenfassend erscheinen vergleichende Studien als zentrales Desiderat für die Big-Data-Forschung (nicht nur) in der Kommunikationswissenschaft, wenn sie theoretischen Mehrwert haben und sich von der reinen Deskription lösen soll. Je mehr Studien publiziert werden, umso eher wird es hoffentlich möglich, auf deskriptiven und häufig auch explorativen Untersuchungen aufzubauen, um durch Big-Data-Studien zu Theoriebildung und -weiterentwicklung beizutragen. Daneben kann auch die Zusammenarbeit mit Kolleginnen und Kollegen aus anderen Fächern den theoretischen Mehrwert von Big-Data-Studien

erhöhen. Ein naheliegender Weg besteht darin, den Kontakt zur Informatik oder Informationswissenschaft zu suchen und damit die eigenen methodischen Kompetenzen und auch technischen Ressourcen zu ergänzen. Dies ermöglicht eventuell komplexere Erhebungs- und Auswertungsverfahren, die ebenfalls über Deskription und Einzelfallstudien hinausgehen. Gerade bei der Arbeit mit ungewohnten Datenmengen und -strukturen, was auf Big Data in der Kommunikationswissenschaft in der Regel zutrifft, können fehlende Kenntnisse und Ressourcen ein zentrales Hemmnis für die Aussagekraft einer Untersuchung sein (boyd & Crawford, 2012; Manovich, 2012).

Dabei kann es aber nicht darum gehen, wie Pfeffer (2013) es in seiner Keynote zuspitzte, sich „einen Informatiker im Nebenzimmer zu halten“. Denn auch der theoretische Austausch mit anderen Disziplinen kann die eigene Arbeit befruchten, sei es durch wirkliche Kollaboration (statt Betrachtung fachfremder Kolleginnen oder Mitarbeiter als reine Dienstleister) oder auch schlicht durch Lektüre der Literatur anderer Fächer. Selbst wenn das Forschungsinteresse fachfremder Untersuchungen von Online-Plattformen vom eigenen Theoriebestand weit entfernt ist, können deren Befunde dennoch Anregungen für die Entwicklung von Theorie (und Methode) einer eigenen Studie liefern. Das Phänomen der Diffusion von Online-Inhalten wird, wenn auch mit anderen Begriffen, beispielsweise auch in der Informatik (z. B. Brodersen, Scellato, & Wattenhofer, 2012; Rodrigues, Benevenuto, Cha, Gummadi, & Almeida, 2011) und der Marketing-Forschung untersucht (z. B. Berger & Milkman, 2012; Nelson-Field, Riebe, & Newstead, 2013). Die Beschäftigung mit Big Data aus Online-Plattformen zeigt damit wiederum auf, dass das interdisziplinäre Fach Kommunikationswissenschaft auch jenseits der Sozialwissenschaften Nachbarn hat, deren Arbeit es lohnt, zur Kenntnis zu nehmen.

Letztlich kann auch die fachinterne methodische Ausbildung dazu beitragen, den theoretischen Ertrag von Big-Data-Untersuchungen zu erhöhen (boyd & Crawford, 2012; Manovich, 2012). Eine Reflexion über Big-Data-Verfahren im Lichte üblicher sozialwissenschaftlicher Gütekriterien erscheint dabei insbesondere dringlich (Mahrt & Scharrow, 2013). In den jüngsten Empfehlungen zur Methodenausbildung der DGPK (Matthes et al., 2011) wird Big Data nicht erwähnt. Und sicherlich wird es in absehbarer Zeit nicht die klassischen sozialwissenschaftlichen Verfahren der Datenerhebung und -analyse ersetzen. Es wäre aber denkbar, an Standorten mit forschungsstarken Studiengängen auch Big-Data-Verfahren in die Methodenlehre aufzunehmen. Im Bereich fortgeschrittener quantitativer Analy-

severfahren gibt es zur Zeit keine festen Standards in der Lehre, und neben oder anstelle von Mehrebenenanalyse oder Strukturgleichungsmodellen – wie von Matthes et al. (2011) vorgeschlagen – könnten auch datenorientierte Verfahren stehen, etwa in Masterstudiengängen mit umfangreichem Methodenanteil. Darüber hinaus könnte die Ausbildung in speziellen Analyseverfahren für große Datenmengen auch zu höherem Anspruch bei der Auswertung von Big-Data-Studien führen. Wie oben dargestellt, sehen Big-Data-Enthusiasten wie Anderson (2008) und Bailenson (2012) Korrelationsanalysen mit sehr großen Datensätzen als vielversprechenden Weg zu Erkenntnis, der Kausalanalysen (zumindest in bestimmten Fällen) ersetzen kann. Viele kommunikationswissenschaftliche Big-Data-Studien führen aber keine Korrelationsanalysen durch, sondern präsentieren streng genommen rein deskriptive und zudem häufig univariate Auswertungen. Mit komplexeren Analysen, die Zusammenhänge zwischen verschiedenen Variablen berücksichtigen, könnte sich auch der theoretische Ertrag einer Studie erhöhen.

Big-Data-Forschung ist nicht per se gut oder schlecht, theoretisch gehaltvoll oder theoriearm. Es liegt an der konkreten Ausgestaltung jeder Untersuchung, ob sie z. B. Problembeschreibung oder Theorienentwicklung leisten will – und kann. Die Banalisierung von theoretischen Konstrukten wird hoffentlich nicht das bestimmende Muster von Big Data in der Kommunikationswissenschaft sein. Welche Rolle der Ansatz langfristig in diesem Fach spielen wird, wird aber voraussichtlich auch und gerade von den theoretischen Impulsen abhängen, die durch Verfahren mit großen Datensätzen geleistet werden.

Dr. Merja Mahrt ist wissenschaftliche Mitarbeiterin am Institut für Sozialwissenschaften an der Heinrich-Heine-Universität Düsseldorf

Quellenverzeichnis

- Adler, P. S., & Kwon, S.-W. (2002). Social capital: Prospects for a new concept. *Academy of Management Review*, 27(1), 17-40. doi: 10.5465/amr.2002.5922314
- Anderson, C. (23. Juni 2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Abgerufen von http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

- Bailenson, J. N. (2012). Beitrag zur Closing Plenary der ICA-Jahrestagung: "The Internet is the end of communication theory as we know it". 62. Jahrestagung der International Communication Association, Phoenix, AZ. Abgerufen von <http://www.icahdq.org/conf/2012/closing.asp>
- Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). *The role of social networks in information diffusion*. Proceedings of the 21st international conference on World Wide Web, Lyon, Frankreich. doi: 10.1145/2187836.2187907
- Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2), 192-205. doi: 10.1509/jmr.10.0353
- Bickham, D. S. (2009). Media effects research as applied science: Social scientists surviving (and prospering) in a newly medicalized field. 59. Jahrestagung der International Communication Association, Chicago, IL. Abgerufen von http://citation.allacademic.com/meta/p297804_index.html
- boyd, d., & Crawford, K. (2012). Critical questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662-679. doi: 10.1080/1369118x.2012.678878
- Brodersen, A., Scellato, S., & Wattenhofer, M. (2012). *YouTube around the world: Geographic popularity of videos*. Proceedings of the 21st international conference on World Wide Web, Lyon, Frankreich. doi: 10.1145/2187836.2187870
- Brooks, B., Welser, H. T., Hogan, B., & Titsworth, S. (2011). Socioeconomic status updates. Family SES and emergent social capital in college student Facebook networks. *Information, Communication & Society*, 14(4), 529-549. doi: 10.1080/1369118x.2011.562221
- Burke, M., Kraut, R., & Marlow, C. (2011). *Social capital on Facebook: Differentiating uses and users*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Kanada. doi: 10.1145/1978942.1979023
- Busemann, K., & Gscheidle, C. (2012). Web 2.0: Habitualisierung der Social Communities. *Media Perspektiven*, (7-8), 380-390.
- Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3), 145-156.
- Dearing, J. W., & Rogers, E. (1996). *Agenda-setting*. Thousand Oaks, CA: Sage.
- Gaffney, D., & Puschmann, C. (2014). Data collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Hrsg.), *Twitter and society* (S. 55-67). New York, NY: Lang.

- Groshek, J., & Groshek, M. C. (2013). Agenda trending: Reciprocity and the predictive capacity of social networking sites in intermedia agenda setting across topics over time. *Media and Communication*, 1(1), 15-27. doi: 10.12924/mac2013.01010015
- Kearon, J., & Harrison, P. (2011). Research robots. A dramatic new way to conduct research & generate insights. Abgerufen von http://www.brainjuicer.com/xtra/BrainJuicer_DigiViduals_Research_Robots_Paper.pdf
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176), 1203-1205. doi: 10.1126/science.1248506
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., ... Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915), 721-723. doi: 10.1126/science.1167742
- Mahrt, M. (2014). Einstellung. In C. Wünsch, H. Schramm, V. Gehrau & H. Bilandzic (Hrsg.), *Handbuch Medienrezeption* (S. 113-127). Baden-Baden: Nomos.
- Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, 57(1), 20-33. doi: 10.1080/08838151.2012.761700
- Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Hrsg.), *Debates in the digital humanities* (S. 460-475). Minneapolis: University of Minnesota Press.
- Markham, A. N., & Buchanan, E. (2012). Ethical decision-making and Internet research: Version 2.0. Recommendations from the AoIR Ethics Working Committee. Abgerufen von <http://www.aoir.org/reports/ethics2.pdf>
- Matthes, J., Kuhlmann, C., Gehrau, V., Jandura, O., Möhring, W., Vogelgesang, J., & Wünsch, C. (2011). Zur Methodenausbildung in kommunikationswissenschaftlichen Bachelor- und Masterstudiengängen. *Publizistik*, 56(4), 461-481. doi: 10.1007/s11616-011-0133-6
- Miller, G. (2011). Social scientists wade into the tweet stream. *Science*, 333(6051), 1814-1815. doi: 10.1126/science.333.6051.1814
- Mitchell, A., & Hitlin, P. (2013). Twitter reaction to events often at odds with overall public opinion. *Pew Research Center*. Abgerufen von <http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion>
- Nelson-Field, K., Riebe, E., & Newstead, K. (2013). The emotions that drive viral video. *Australasian Marketing Journal*, 21(4), 205-211. doi: 10.1016/j.ausmj.2013.07.003

- Pfeffer, J. (2013). *Big data, big research? Opportunities and constraints for computer supported social science*. Keynote zur „Digital methods“-Tagung der DGPUK-Fachgruppe Computervermittelte Kommunikation, Wien. Abgerufen von <http://www.pfeffer.at/slides/DigitalMethods-BigData.pdf>
- Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K., & Almeida, V. (2011). *On word-of-mouth based discovery of the web*. Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, Berlin. doi: 10.1145/2068816.2068852
- Russom, P. (2011). Big data analytics. Abgerufen von http://www.cloudtalk.it/wp-content/uploads/2012/03/1_17959_TDWIBigDataAnalytics.pdf
- Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, 41(5), 885-899. doi: 10.1177/0038038507080443
- Scharkow, M., & Vogelgesang, J. (2011). Measuring the public agenda using search engine queries. *International Journal of Public Opinion Research*, 23(1), 104-113. doi: 10.1093/ijpor/edq048
- Schnell, R., Hill, P. B., & Esser, E. (2005). *Methoden der empirischen Sozialforschung* (7. Aufl.). München: Oldenbourg.
- Tinati, R., Halford, S., Carr, L., & Pope, C. (2014). Big data: Methodological challenges and approaches for sociological analysis. *Sociology*, 48(4), 663-681. doi: 10.1177/0038038513511561
- Verdegem, P., & D'heer, E. (2013). *The role of Twitter in political agenda-setting during elections: Findings from the 2012 elections in Belgium*. Vortrag auf der Konferenz „Twitter and Microblogging: Political, Professional and Personal Practices“, Lancaster, UK. Abgerufen von <http://hdl.handle.net/1854/LU-3200218>
- Watts, Duncan J., & Dodds, Peter S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4), 441-458. doi: 10.1086/518527
- Williams, D. (2006). On and off the 'Net: Scales for social capital in an online era. *Journal of Computer-Mediated Communication*, 11(2), 593-628. doi: 10.1111/j.1083-6101.2006.00029.x
- Yoder, C., & Stutzmann, F. (2011). *Identifying social capital in the Facebook interface*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Kanada. doi: 10.1145/1978942.1979025
- Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, 12(4), 313-325. doi: 10.1007/s10676-010-9227-5